Contract 68HERC19D0011
Task Order: # 68HERC20F0025
Task Order Title: Technical Expertise Support on Environmental Protection Agency (EPA) Scientific Issues and Topics

Title: Advances Made During Application of Artificial Intelligence and Open Data Practices in Chemical Hazard Assessment
Anticipated Workshop Date: FY21 Q3
Report Output: Proceedings-in-Brief

Background/Summary: Systematic review (SR) methods are a natural fit for literature-based chemical assessments, providing a mechanism for meeting the expectation that an assessment will include the most relevant and robust information for chemical hazard identification. However, practical application of SR methodology is heavily constrained by the labor intensive and costly nature of deploying SR workflows that include finding all relevant studies (information retrieval), extracting methods and findings, analyzing that information, and making extracted information reusable and interoperable across SR applications. These constraints are amplified in that assessments are often conducted manually, on an as-needed basis by independent entities often unaware of each other's activities leading to significant costs and potential for errors further amplified by the same task being repeated multiple times (and not uniformly). Advances in artificial intelligence (AI) hold promise to bypass these constraints through semi-automation of SR workflows. However, computationally intelligent approaches toward information retrieval are hindered by a lack of high-quality training, test, and validation data.

A related topic is the management of data extracted during SR workflows. Study methods and findings are almost exclusively recorded using highly variable natural language. This presents a significant semantic challenge because linguistic variation can obscure concepts and relationships needed for information retrieval and interpretation. In order to maximize the potential of AI and the benefits established by the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles the environmental health field should explore the use of controlled vocabularies and ontologies in information digitization.

Systematic review methodology is conducted in a step-wise structured manner that is amenable to technological innovation. Therefore, the overall aim of this workshop is to explore strategies to address several advances and lingering challenges faced during application of AI in chemical hazard assessment.

Topics:

- **Increasing efficiency in data extraction pipelines using systematic methods**
  - Present state of data extraction pipelines in systematic review methods
  - Present areas of investigation to create data extraction efficiencies including natural language processing (NLP)/machine learning (ML) methods (referred to generically as artificial intelligence or AI) and tool interoperability
  - Discuss approaches to increase availability of training data for AI
  - Evaluate the accuracy and efficiency of AI technologies for data extraction
  - Controlled vocabularies and ontologies for data management. Present examples of implementation of controlled vocabularies and ontologies for data management
  - Efforts towards standardized data collection by journals/reporting by authors using schemas that support downstream information processing and dissemination
- **Identify cross-cutting issues, e.g., need for interoperability across software applications**

Contract 68HERC19D0011
Task Order: # 68HERC20F0025
Task Order Title: Technical Expertise Support on Environmental Protection Agency (EPA) Scientific Issues and Topics

Plenary Overview: EPA staff will make some presentations on the topics 1-2 highlighted below. We'd like to have non-EPA presenters as well and can provide some suggested names. A total of ~10 plenary presentations across 2 days is envisioned. We would also like a written report. For topic 3 we would like to have break-out rooms where participants can engage with demonstrators of various tools.

1. **Increasing efficiency in data extraction pipelines in systematic methods**
   a. Overview of EPA and Federal Partner's development and implementation of more efficient data extraction pipelines in chemical assessment workflows (Kris Thayer)
      i. Presentations
         1. Full text annotation and Information Extraction in EPA systematic review human health assessment workflows: Michele Taylor
         2. Quantitative data extraction, dose response analysis, and visualization: Andy Shapiro
         3. Automated information retrieval using Linear/Pathway based SR Workflows: Kristan Markey
         4. Utilizing automated and semi-automated data analytic tools for curating data in the ECOTOX Knowledgebase: Dale Hoff
         5. Stewardship, transparency, cross-validation, and trust in AI/semi-automated approaches: Malcom MacLeod

2. **Controlled Vocabularies and Ontologies for Data Management**
   a. Overview of ORDs implementation of controlled vocabularies and ontologies in chemical assessments (George Woodall)
      i. Demonstrations
         1. Data management using controlled vocabularies: Michelle Angrish
         2. Connecting health effects data using UMLS: Sean Watford
         3. Information management in the Comparative Toxicogenomics Database (CTD): Carolyn Mattingly
         4. Information management in the NTP Chemical Effects and Biological Systems (CEBS) Database: Jennifer Fostel or Charles Schmidt
         5. Standardization of data entries submitted by authors to journals: Paul Whaley
         6. Environmental Health Language Collaborative: Stephanie Holmgren

3. **Tool Demo and Poster Session[1] (added footnote with SR Toolbox URL: http://systematicreviewtools.com/):**
   a. Tool demo and poster abstract submission open to the public.
   b. Submission period will be open for 30 days.

[1] http://systematicreviewtools.com/

c. Submissions will be reviewed and selected by the organizing committee.

d. Posters: Seeking 10-15 external poster presentations (to supplement posters prepared by EPA staff, of up to 20-25 posters). EPA will provide abstracts and posters directly to the NAS in advance of the workshop.

e. Tool Demonstrations

i. Demos for information extraction tools

1. Fiddle (Sciome)
2. PDF ANotation and Data HArmonization (PANDHA; ORNL/EPA)
3. [HYPERLINK "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5801564/"] (Nancy Baker, CCTE)

ii. Demos for Systematic Review Workflows

1. HERO
2. DistillerSR
3. HAWC

iii. Demos for ontology resources

1. Biocreative
2. Bioportal
3. UMLS (Pertti Bert Hakkinen)

- **Abstract submission guidelines (see attached abstract template)**
  - Must include title, authors, affiliations, and COI.
  - Should indicate the specific objective that will be addressed in the submission
  - 300-word limit (applies only to abstract body).